Dealing with Distributional Assumptions in Preregistered Research

Williams, Matt N.; Albers, Casper J.

Published in:
Meta-Psychology

DOI:
10.15626/MP.2018.1592

Meta-Psychology, 2019, vol 3, MP.2019.1592,
https://doi.org/10.15626/MP.2019.1592
Article type: Original Article

Open data: Yes
Open materials: Not relevant
Open and reproducible analysis: Yes
Open reviews and editorial process: Yes
Preregistration: Not relevant

Edited by: Moritz Heene
Reviewed by: Felix Naumann, Sven Hilbert
Analysis reproduced by: Rickard Carlsson
All supplementary files can be accessed at the OSF project page: https://doi.org/10.17605/OSF.IO/DZN3S

Dealing with Distributional Assumptions in Preregistered Research

Matt N. Williams
Massey University, New Zealand

Casper J. Albers
University of Groningen, Netherlands

Virtually any inferential statistical analysis relies on distributional assumptions of some kind. The violation of distributional assumptions can result in consequences ranging from small changes to error rates through to substantially biased estimates and parameters fundamentally losing their intended interpretations. Conventionally, researchers have conducted assumption checks after collecting data, and then changed the primary analysis technique if violations of distributional assumptions are observed. An approach to dealing with distributional assumptions that requires decisions to be made contingent on observed data is problematic, however, in *preregistered* research, where researchers attempt to specify all important analysis decisions prior to collecting data. Limited methodological advice is currently available regarding how to deal with the prospect of distributional assumption violations in preregistered research. In this article, we examine several strategies that researchers could use in preregistrations to reduce the potential impact of distributional assumption violations. We suggest that pre-emptively selecting analysis methods that are as robust as possible to assumption violations, performing planned robustness analyses, and/or supplementing preregistered confirmatory analyses with exploratory checks of distributional assumptions may all be useful strategies. On the other hand, we suggest that prespecifying "decision trees" for selecting data analysis methods based on the distributional characteristics of the data may not be practical in most situations.

Keywords: preregistrations, distributional assumptions, open science, transparency.

We thank Moritz Heene as editor and reviewers Felix Naumann and Sven Hilbert for valuable feedback on an earlier version of this article. We also thank Tobias Mühlmeister for assisting in the copyediting of our manuscript.

Correspondence regarding this article can be addressed to Matt N. Williams, School of Psychology, Massey University, Private Bag 102903 North Shore, Auckland 0745, New Zealand. Email: M.N.Williams@massey.ac.nz

Virtually any inferential statistical method relies on some set of distributional assumptions[1] in order to produce valid inferences. For example, a regression model estimated via ordinary least squares relies on the assumptions the predictors are measured without error, that any measurement error in the response variable is uncorrelated with the predictors, and that the error terms are independent, identically and normally distributed with a mean of zero for all values of the predictors[2] (Williams, Grajales, & Kurkiewicz, 2013). Even non-parametric methods have assumptions, albeit not with respect to the specific probability distribution of particular variables. For example, if a Mann-Whitney test is used to test the equality of the medians of two populations on some variable, one must assume that the distribution of the variable has the same shape and spread within each of the populations (Fagerland & Sandvik, 2009). Distributional assumptions are a common source of misconceptions (Ernst & Albers, 2017), but even when assumptions are correctly identified and investigated, several issues can arise. One of these will be discussed in this paper.

Breaches of distributional assumptions can cause problems for inference, including biased estimates, artificially narrow (or broad) confidence intervals, and increases in Type I and/or Type II error rates (Ernst & Albers, 2017; Williams et al., 2013). The severity of these problems varies depending on the analysis method, the sample size, the nature of the assumption breach, and on whether one or more assumptions are violated simultaneously. The consequences of an assumption breach can vary from a minor change in Type I error rates and confidence interval coverage, through to biased parameter estimates, right through to parameter estimates fundamentally losing their intended interpretation. For example, in a simple linear regression model estimated using ordinary least squares, a breach of the assumption that the error terms are normally distributed will not cause biased or inconsistent estimates or harm the interpretability of the parameters, but only affect confidence interval coverage and Type I error rates, and even these effects can be mild (see Gelman & Hill, 2007; Lumley, Diehr, Emerson, & Chen, 2002; Meuleman, Loosveldt, & Emonds, 2015; Williams et al., 2013). On the other hand, if the assumption of a linear relationship between the predictor variable and the response variable is breached, the slope loses its interpretability as a measure of the dependency between the predictor and response variables (see Meuleman et al., 2015): A measure of linear relationship is of little value if the true relationship is not linear.

Many methodological textbooks and other resources offer researchers advice on how to detect and respond to distributional assumption violations. There are many methods for detecting distributional assumption problems, including both graphical approaches and inferential tests. For example, a researcher interested in whether the error terms in her regression model are normally distributed could evaluate this assumption using visual methods, such as a q-q plot, a formal statistical test such as the Shapiro-Wilk or Kolmogorov-Smirnov test, or by evaluating skewness and kurtosis statistics. Likewise, the potential responses available for dealing with a distributional problem are legion, including transformations, deletion of outliers, trimming of samples, alternative estimation algorithms, randomisation-based tests, rank-based non-parametric statistics, and many others.

[1] By "distributional assumption" we mean an assumption with respect to the univariate, bivariate, or multivariate distribution of variables and/or error terms—e.g., that the relationship between two variables is linear, or that the variances of a set of error terms is identical, or that the distribution of a response variable is negative binomial conditional on a set of predictor values. Such assumptions are also sometimes referred to as "statistical assumptions" or just "assumptions". We use the modifier "distributional" simply to differentiate such assumptions from purely non-statistical assumptions (e.g., assumptions about ontology or epistemology).

[2] The assumption that the error terms all have mean zero for any combination of values of the predictor variables implies that the independent effects of the predictor variables included in the model on the response variable are *additive* and *linear*. Indeed, some presentations of the assumptions of multiple regression (e.g., Gelman & Hill, 2007) replace a description of this assumption with a description of the assumption of "additivity and linearity". Note that the predictor variables included in a regression model may include transformations of the original variables (e.g., polynomial terms), which provides some capacity to specify nonlinear effects between the original variables in a dataset and the response variable.

Nevertheless, an important *meta*-strategy underlies the advice about dealing with distributional assumptions found in many methodological resources (see the discussion in Wells & Hintze, 2007): First, one should check for distributional problems, and then, *if* problems are detected, select a strategy to deal with the problems. We term this the "test then respond" meta-strategy for dealing with distributional assumption violations.

The potential risks of the "test then respond" meta-strategy will be apparent to readers aware of the problems of the "garden of forking paths" (Gelman & Loken, 2014, p. 460) and "researcher degrees of freedom" (Simmons, Nelson, & Simonsohn, 2011, p. 1359). By applying different analysis strategies contingent on the characteristics of the observed data, a researcher may end up happening upon a statistically significant result in favour of a particular hypothesis that is itself contingent on an analysis decision made after observing the data. This will be especially problematic if the researcher is motivated to search for and selectively report statistically significant results ("*p*-hacking"). *p*-hacking can result in inflated Type I error rates and seriously biased estimates, and is thought to be one of the major causes of the current "replication crisis" in psychology and other sciences.

Although *p*-hacking has especially negative effects, the making of analysis decisions contingent on observed data can still be problematic even if the researcher's intentions are entirely scrupulous. Specifically, the nominal error rates of analysis strategies are invariably derived based on repeated sampling with a fixed analysis method, and will not necessarily apply where the analysis method depends on the data. For example, the common strategy for comparing two means of using a Student's *t* test if a preliminary Levene's test fails to reject the null hypothesis of equal variances across the two groups, but using a Welch's *t* test if the Levene's test is significant, can result in Type I error rates that differ markedly from the nominal alpha level (Albers, Boon, & Kallenberg, 2000; Bancroft, 1964; Zimmerman, 2004).

Preregistration

One important strategy gaining popularity as a partial solution to problems with replicability and *p*-hacking is *preregistration*. In a landmark paper, Wagenmakers, Wetzels, Borsboom, van der Maas and Kievit (2012) argued that when research is intended to be confirmatory—i.e., to test hypotheses—a data collection and analysis plan should be preregistered in advance. Doing so reduces the capacity of researchers to exploit flexibility in their data collection and analysis procedures to produce positive (or statistically significant) findings. Preregistration is a crucial control strategy in an environment where researchers are incentivised to produce statistically significant and novel findings in order to achieve publications in high-impact journals. Online platforms for uploading and permanently time-stamping preregistrations have since been developed (osf.io, aspredicted.org), and preregistered research studies have been increasingly frequent in the pages of psychology journals, especially within experimental social psychology; see for example the 67(1) special issue of the *Journal of Experimental Social Psychology*, which was entirely dedicated to preregistered research.

Preregistration is useful for increasing the credibility of individual studies, but it can also help address the broader problem of *publication bias*—a problem wherein statistically significant findings are more likely to be published than non-significant ones (see Ferguson & Heene, 2012). Publication bias can distort the mean effect sizes estimated in meta-analyses both because studies that produce non-significant findings are less likely to end up in the published literature, and because authors may respond to the existence of this bias by exploiting flexibility in data collection and analysis procedures to produce statistically significant findings. Preregistration can help to address publication bias both in the sense that the individual preregistered studies included in a meta-analysis may be less likely to have used methods that produce biased effect sizes (e.g., *p*-hacking), but also in the sense that researchers conducting meta-analyses can search for preregistrations that have *not* resulted in published outputs, and thus obtain some

information about the size of the unpublished "file drawer".

Preregistration is clearly valuable, but it is currently unclear how researchers should deal with distributional assumptions when performing preregistered research. The conventional "test then respond" approach to dealing with distributional assumptions—where the researcher selects a primary analysis strategy, performs exploratory assumption checks using a range of statistical and graphical measures, and then uses their judgment to determine whether a change in analysis strategy is needed—is clearly anathema to a preregistered approach to research. So how should researchers writing preregistrations deal with distributional assumptions? In this article we aim to provide concrete and practical advice to help researchers pre-emptively respond to the spectre of breaches of distributional assumptions when performing preregistered research.

Strategies for Dealing with Distributional Assumptions in Preregistrations

One strategy for dealing with distributional assumptions in preregistered research is to simply ignore the issue and preregister a primary analysis strategy without any conscious attention to distributional assumptions whatsoever. This strategy does at least avoid the possibility of biased estimates caused by the researcher exploiting analytic flexibility to produce statistically significant findings. In this sense it may arguably be superior to a strategy of using exploratory strategies to diagnose distributional problems, running multiple analyses, and potentially allowing decisions about which analyses to report to be affected by whether or not they produce significant results.

This said, selective reporting or p-hacking is obviously not the only cause of biased estimates or untrustworthy findings. Some distributional assumption violations can cause very serious inferential problems (consider, for example, the bias in estimates that can result when predictors in regression models are measured with error; Westfall & Yarkoni, 2016). As such, simply ignoring distributional assumptions in preregistered research is obviously not a complete solution. Alternatively, there are a variety of more sophisticated approaches that a researcher could make use of for dealing with distributional assumptions in the context of preregistered research. In the current section, we consider and discuss four potential strategies in turn. In discussing each strategy we will focus on how these strategies can be conducted and their consequences for the resulting analyses of empirical data, but we will also briefly touch on the implications of these strategies for analyses conducting *before* data collection—i.e., statistical power and sample size determination.

Strategy 1: Decision Tree

The first strategy is to outline a decision tree specifying what methods will be used to identify distributional assumption breaches, and which alternative methods will be applied if breaches are identified. For example, a researcher might specify that linear regression using ordinary least squares will be applied in order to test their hypothesis, and a Shapiro-Wilk test applied to the residuals. If the Shapiro-Wilk test statistic is *not* significant, confidence intervals will be calculated based via the usual Wald method; if it *is* significant (indicating non-normality), they will be calculated using a percentile bootstrap. Such a decision tree is effectively just a preregistered form of the test then respond meta-strategy, albeit with a commitment to pre-specified criteria for making decisions. The decision tree method is currently implicitly endorsed in the widely used template for preregistrations in social psychology (van 't Veer & Giner-Sorolla, 2016), which asks researchers to specify "Assumptions of analyses, and plans for alternative/corrected analyses if each assumption is violated." However, we suggest that this strategy does have several problems.

Problem 1: Uncertainty involved in diagnosing distributional problems. The first problem is that determining whether a particular distributional assumption violation is present—and of a magnitude likely to harm inferences—can be difficult, and the conclusion of such an investigation may come with great uncertainty attached. Consider, for example, a researcher conducting a simple experiment with a continuous response variable and two conditions (treatment and control). Here, the primary research question might be simple: Is the mean of the response variable higher for treated participants

than control participants? The most obvious analysis strategy would also be simple: An independent-samples Student's *t* test. Yet the questions to be investigated in an examination of the distributional assumptions pertaining to this test would be much more complex.

For example, investigating the assumption of independent error terms means answering this question: Over repeated sampling, would each of the N error distributions in this design be statistically independent of one another, and, if not, is the form and magnitude of this non-independence sufficient to cause substantive changes to the error rates of the *t* test in the main analysis? It should be obvious that this distributional question is more complex than the primary research question regarding the differences between two means. It is complex both because it refers to the relationships between a large *matrix* of error terms (rather than just two means), about which we must make inferences based on a single sample of residuals, and also because it is *double-barrelled*—the question is not just whether the error terms are statistically dependent, but whether this dependence in turn is likely to harmfully affect the resulting inferences in the main analysis. The net result would be that any investigation of the validity of this assumption is likely to come with substantial uncertainty attached, and as such relying on such an investigation to determine the final choice of primary analysis may be problematic.

Problem 2: Difficulty of formulating hard-and-fast rules. Relatedly, it is often difficult to formulate hard-and-fast rules for diagnosing particular distributional assumption breaches. Many conventional distributional assumption diagnostics require the researcher to subjectively interpret a plot (e.g., a q-q plot for diagnosing non-normality, an autocorrelation function plot for diagnosing correlated error terms, a scatterplot of predicted values vs. residuals for diagnosing heteroscedasticity or non-linearity, etc.). Using such graphical methods to detect assumption violations is explicitly recommended by the APA's statistical task force (Wilkinson & Task Force on Statistical Inference, 1999). In the context of preregistration, however, which is largely intended to *remove* flexibility in how data analysis is conducted, relying on the researcher to subjectively interpret a plot and then make a decision is problematic.

Firmer statistical rules for diagnosing particular problems exist, of course. For example, one can statistically test a null hypothesis that a particular distributional assumption is valid. A Shapiro-Wilk test, for instance, can be used to test a null hypothesis that the error terms for a particular model are normally distributed. There is a paradox here, though, in the sense that statistical tests will generally have low power to detect distributional problems when the sample size is small, even though this is precisely the scenario in which distributional assumptions matter most. On the other hand, if the sample size is large, statistical tests of assumptions may be powerful, even though the large sample size means that the primary analysis may be robust to the detected assumption violation[3]. Of course, this is a general problem with statistical tests of assumptions, rather than one that solely applies in the context of preregistration or the decision tree strategy.

Problem 3: Complexity of decision trees required. The third problem is the fact that, even if it were possible to accurately and objectively diagnose distributional problems, there is a large number of distributional problems that may arise for any given analysis, and thus an exploding quantity of potential remedies. For example, a simple linear regression model can be afflicted by a wide variety of distributional problems, including measurement error of any sort in the predictor variable, measurement error in the response variable that is correlated with the predictor, non-linearity of the relationship between the predictor and the response variable, dependent error terms, heteroscedasticity of errors, or non-normality of

[3] This admittedly depends on the nature of the assumption violation. For example, a linear regression model estimated using ordinary least squares will be more and more robust to a violation of the assumption of normally distributed error terms as the sample size increases, due to the central limit theorem (provided that the other assumptions of the model hold). On the other hand, if measurement error is present in the predictor variables, increasing the sample size would not necessarily reduce the biasing effect of this measurement error on the estimates of the regression parameters.

errors. Often these problems occur in combination with one another. Each problem has an array of potential remedies: For example, heteroscedasticity might be dealt with by using Huber-White "sandwich" errors, by variance-stabilising transformations, or by bootstrapping (see Liu, 1988). Further checks may be necessary to check whether particular remedies "worked", those checks potentially implying the need for yet more decisions (did the variance stabilising transformation produce homoscedasticity? If not, what potential remedy should be tried next?) Furthermore, the order in which the assumptions are checked—although arbitrary—can have an effect on the final model that is applied to the data. Setting out a decision tree that will select an appropriate analysis strategy for each of the various combinations of problems that may occur for a given analysis would thus be an extremely difficult task, and the prospect of doing so might discourage researchers considering using preregistration.

Problem 4: Effects on nominal error rates. A final problem with the "decision tree" approach to diagnosing and responding to distributional assumption problems in preregistrations is the fact that, whatever the primary analysis technique ends up being, its nominal Type I and Type II error rates (e.g., the pre-set alpha and 1-power) will typically be based on an assumption that the analysis technique was fixed in advance. Unless they perform simulation studies, researchers will not typically know what the applicable error rates are in the scenario of analysis decisions being made contingent on particular features of the data. As noted previously, these error rates may vary considerably from the nominal error rates of the individual techniques considered in isolation (e.g., Zimmerman, 2004).

Specifying simulations to estimate error rates and power for the decision tree approach could often present significant challenges. Such an analysis would require a simulation in which multiple samples are simulated from a population in which there is a particular hypothesised effect size, the analysis method for each sample is decided according to the preregistered decision rule, and the data analysis then conducted. This alone may be challenging for many researchers to program, but an even more difficult challenge would be deciding what distributional characteristics or misspecifications to incorporate in the simulated data. The researcher is thus faced with the prospect of trying to predict the types of distributional problems that are likely to occur and to then simulate data that embodies these problems, which will inevitably require some complex (and relatively arbitrary) decision-making.

Strategy 1: Conclusion. In summary, while the decision tree approach to dealing with distributional assumption breaches in preregistrations may be appropriate when wielded by expert researchers in some circumstances, we suggest that it has several important problems that mean that it is not suitable as a default strategy in preregistrations.

Strategy 2: Selecting a Robust Primary Analysis Strategy

Given the need to make analysis decisions prior to any opportunity to check for distributional assumption problems, it may be useful to select primary analysis strategies that are as *conservative* as possible with respect to those assumptions. As stated earlier, virtually any inferential analysis will require distributional assumptions of some sort, but some analyses require stronger assumptions than others. Some examples of analysis choices that may reduce the reliance on at least some distributional assumptions in commonly used statistical analyses include:

- Using bootstrapping or permutation tests rather than normal theory to calculate confidence intervals and p values (thus obviating the need for normally distributed errors in linear models)
- Using Huber-White "sandwich" standard errors rather than assuming homoscedasticity in regression-like models (White, 1980)
- Using so-called robust methods that are designed to rely less on distributional assumptions (Wilcox, 2012).

Beyond these familiar examples, the possibility of using Bayesian models estimated using Markov Chain Monte Carlo (MCMC) means that researchers can quite readily estimate models where specific assumptions are loosened in specific ways. For example, a Bayesian regression model can be estimated in which the distribution of the error terms is modelled not with a normal distribution, but rather, for example, a skew-normal distribution

in which a skewness parameter is freely estimated (and in which a the normal distribution is a special case; see Azzalini, 1985). Similarly, one can specify a Bayesian regression model in which the variance of the errors is not necessarily constant across all the values of the predictor variables, but instead some function of the values of the predictor variables (with constant variance again being a special case). Bayesian models are by no means assumption-free, but do give the researcher the capacity to thoughtfully loosen specific distributional assumptions, and thus may be an attractive option in the context of preregistered research. Excellent introductions to Bayesian data analysis can be found in Kruschke and Liddell (2018) and Etz and Vandekerckhove (2018). The programming language Stan (Carpenter et al., 2017) provides a framework for implementing Bayesian analyses that apply flexible sets of assumptions.

In attempting to select a robust primary analysis method, researchers should carefully consider what distributional issues or problems are most likely to arise in the context of their specific study. A researcher conducting a study will often have some familiarity with the measurement instruments being used (whether these are survey scales, electromyography devices, counts of particular incidents, or reaction times), and the typical characteristics of the data produced by these instruments. Researchers may thus be able to draw on their own contextual knowledge to anticipate the distributional issues that might arise, and pick a statistical analysis method that is robust to the most plausible problems.

The approach of selecting a primary analysis method that makes distributional assumptions that are as weak as possible can definitely be useful in preregistrations. Doing so avoids the need to make analytic decisions contingent on data. In comparison to the other strategies discussed in this section, this strategy also minimises the need for conducting multiple data analyses, thus streamlining the analysis process and resulting in a more concise write-up. This said, this strategy can also be applied in combination with the two strategies we will consider below.

However, this strategy is not without its problems either. If all the assumptions of the linear model are met, then the standard parametric methods are uniformly most powerful. The benefit of not having distorted Type I error rates if the assumptions are violated comes at the cost of structurally lower power, thus a higher Type II error rate, if the assumptions actually hold true. Admittedly, the extent of this problem differs depending on the planned test. For instance, the Welch *t* test, for unequal variances, barely has lower power than the standard Student *t* test (Delacre, Lakens, & Leys, 2017).

Power analysis may again be challenging when applying the robust primary analysis strategy, mainly in that there may be ambiguity in terms of the distributional characteristics that should be assumed when simulating data for the purposes of power analysis. For example, if a Welch's *t* test is planned, should we nevertheless simulate data from two populations with equal variances? If not, how different should the variances be? This said, the complication of programming a simulation in which the analysis method applied to each sample differs depending on its characteristics (as for Strategy 1) is avoided, making power analysis slightly easier for Strategy 2 (robust primary analysis strategy) than is the case for Strategy 1 (decision tree).

Strategy 3: Robustness Analysis

The third strategy we consider here is that of deliberately preregistering *multiple* analyses that answer the same research questions (while making different distributional assumptions). We will term this strategy that of performing *robustness analyses.* Two related terms are *sensitivity analysis* (which investigates how uncertainty pertaining to the inputs into scientific models relate to uncertainty in the outputs; see Saltelli, Tarantola, Campolongo, & Ratto, 2004), and *multiverse analysis* (which focuses on how different decisions made during data processing can lead to different datasets and different conclusions; see Steegen, Tuerlinckx, Gelman, & Vanpaemel, 2016). We prefer the term of robustness *analysis* over robustness *checks* because we see the purpose of this strategy as to investigate how the results vary across numerous plausible choices of analysis, rather than just to check that some favoured conclusions holds up under an alternative specification.

As an example of the robustness analysis approach, imagine a researcher seeking to test the hypothesis that one continuous variable has a

positive relationship with another continuous variable (with both variables comprising series of observations gathered over time). Here the researcher might preregister a simple ordinary least squares regression as a primary analysis, with a model that also includes an autoregressive correlated error structure as a robustness analysis. This robustness analysis would help to deal with the potential problem of a breach of the assumption of independent error terms, which is a problem that often arises when analysing time series data. Robustness analyses are useful especially when there is genuine ambiguity over which analysis technique is the most appropriate choice to address a given research question.

When employing this strategy, one might designate one particular analysis as the primary analysis, and a set of other analyses as the robustness analyses. One could also regard *all* the planned analyses as having equal priority (and all being "robustness analyses"). Regardless, the key to an effective preregistered robustness analysis is identifying different analysis methods that test the *same* (preregistered) hypotheses, but while making different (but plausible) distributional assumptions. As some examples:

- If planning a correlational analysis, one could specify both an analysis that assumes that the variables have a linear relationship (Pearson's correlation) as well as an analysis that assumes only a monotonic relationship (Spearman's rho).
- If planning a linear regression, one could specify an analysis that assumes homoscedasticity (OLS estimation), as well as an analysis that produces consistent estimates even in the presence of heteroscedasticity (Huber-White sandwich standard errors).
- If planning a linear regression, one could specify both an analysis using standard OLS estimation as well as a model including an autoregressive AR(1) term, to allow for possible serial dependence in the data.

Synthesising the results from robustness analyses. It is possible to preregister a specific decision rule for interpreting the results of multiple analyses (e.g., "If the relationship is positive and statistically significant in both the Pearson's correlation analysis and the Spearman's rho analysis, we will conclude that the data supports the hypothesis"). However, such decision rules are necessarily arbitrary, and designing a sensible interpretation structure may be more difficult for a larger number of analyses (e.g., what if the coefficient is statistically significant in six out of seven planned robustness analyses?) There is also no strong reason to assume that a combined hypothesis test based on multiple statistical methods would have more desirable long-run properties than any one of the single analyses that are included within the combined test. As such, it may be more appropriate to preregister multiple robustness analyses, preregister criteria for the interpretation of each, and to subsequently attempt to sensibly integrate the findings produced across these analyses—but not to specify an overarching decision rule based on the combined results of multiple analyses. This means that the synthesis of findings (e.g., in a discussion section) may be more complex when applying robustness analyses than when using any of the other strategies, with more ambiguity about whether a particular set of findings supports or does not support a particular hypothesis. It may nevertheless allow the researcher to clearly communicate the degree to which a particular finding "holds up" across multiple reasonable analysis options.

Power in robustness analyses. Conducting a power analysis when planning to apply Strategy 3 could be either straightforward or very complex. If a relatively simple analysis method with strong assumptions is selected as the primary analysis strategy (in conjunction with some additional robustness analyses with different assumptions), it could be justifiable to perform the power analysis solely for the primary analysis strategy. This means that it could be possible to perform the power analysis using point-and-click software such as G*Power (Faul, Erdfelder, Buchner, & Lang, 2009). Such an analysis would nevertheless need to come with an acknowledgement that the power of the additional analyses is almost inevitably likely to be lesser (depending in part on the actual degree to which any assumptions are breached), and that the reported power analysis should be interpreted as representing a best-case scenario. When taking this approach, it would probably be helpful to increase the planned sample size somewhat beyond what the power analysis suggests is required.

On the other hand, a more comprehensive power analysis would check power for all of the planned robustness analyses, and do so based on data that is simulated so as to display plausible assumption violations. Such a power analysis could be challenging to conduct.

Robustness analysis and meta-analyses. A problem with the robustness analysis approach is the ambiguity in terms of how the findings of a study that conducted a robustness analysis should be coded if included a later meta-analysis, given that each statistical method included in the robustness analysis may have produced a different estimated effect size. A researcher conducting a meta-analysis and aiming to code the effect sizes reported in an individual study conducted using robustness check may thus face uncertainty about whether to include just one effect size estimate (and if so, which one), or whether to aggregate the findings of the various analyses in some way. While this may be an ambiguity that can satisfactorily be resolved within the preregistration for a given meta-analysis (see Quintana, 2015), the use of a robustness analysis within an individual study probably presents greater complications for meta-analysis than do the other strategies discussed in this article.

Strategy 4: Exploratory Assumption Checks

The final strategy we consider here is preregistering a primary (confirmatory) analysis method that will be followed regardless of the characteristics of the data, without necessarily making this primary analysis method a robust one, but including in the ensuing data analyses some *exploratory* investigations of distributional assumptions (or "assumption checks"). A plan for these investigations could be specified to at least some degree in the preregistration (perhaps making them more confirmatory in nature), but this is not absolutely necessary—provided that the analyses are explicitly tagged as exploratory in the final report.

This strategy has the advantage of allowing for the communication of information about distributional assumptions without increasing the complexity of the preregistration too greatly. It also makes power analysis straightforward, since there would be just one primary analysis method to plan (for each research question), and the analysis method might well be a commonly-used one for which power analysis is available in easy-to-use software such as G*Power. Furthermore, the fact that the distributional assumption checks would not need to be used to make binary decisions (unlike the case in strategy 1) means that graphical methods could be used to convey information about the magnitude of any assumption breaches.

The primary downside of this strategy would be the resulting ambiguity with respect to how the findings of the assumption checks should impact the interpretation of the results of the primary analysis. We would essentially suggest that the preregistered primary analysis should be conducted, reported and interpreted as planned in the preregistration effectively regardless of the outcomes of the distributional assumption checks, but that the researcher should identify any apparent distributional problems as a reason to interpret the results with some extra caution. There is a risk here, however, that a researcher might describe and emphasise evidence for a particular distributional problem differently depending on whether the results of the primary analysis are "positive" or not (e.g., ignoring a distributional problem if the main findings are positive, vs. emphasising the distributional problem as a possible explanation of the results if the data would otherwise appear to falsify a favoured theory). As such, this strategy has its dangers, but may be useful in some contexts—particularly student research, where preregistering a relatively simple analysis plan and using a relatively simple power analysis, while allowing some capacity for distributional assumption checks, may be desirable.

Example

In this section, we will give a practical example of the four strategies suggested in this paper. In a paper published in Psychological Science, Schroeder and Epley (2015) reported multiple related organizational psychological studies. These studies investigated the effect of voice on (hypothetical) job applications. In this paper, we look only at their Experiment 4, and use the description of this experiment by McIntyre (2016).

In this experiment, 39 professional recruiters were assigned to one of two conditions. The recruiters either listened to a spoken job application

or read a written transcript of the application. The recruiters rated the participants on intelligence, competence, and thoughtfulness, resulting in an average rating denoted as intellect. Some covariates were measured as well, but we will ignore these for the sake of simplicity. The main research question was whether ratings differed between spoken and written job pitches. Even for such a seemingly simple design, there are many researcher degrees of freedom. For instance, the choice to define "intellect" as the arithmetic mean of competence, thoughtfulness and intelligence implies that these three variables are equally important ingredients of intellect.

.

Table 1. *Data for Experiment 4 of Schroeder & Epley (2015)*

Written job pitch	1.67	2.00	2.67	2.67
	3.00	3.33	4.33	4.33
	4.67	4.67	4.67	4.67
	5.67	5.67	6.67	7.33
	7.67	8.00		
Spoken job pitch	3.33	4.33	4.67	5.67
	5.67	6.00	6.00	6.00
	6.33	6.67	6.67	6.67
	7.00	7.00	7.00	7.00
	7.00	7.67	8.67	10.00
	10.00			

As these issues are besides the focus of our paper, we will not discuss them further, and instead take the 39 intellect ratings as given and work from there. The scores are given in Table 1, and are also available (along with analysis code) in the OSF project for this paper https://osf.io/h2xry/ . Below, we indicate how the test for comparing both experimental groups could be preregistered, according to the four strategies. The beginning of the preregistration, describing the data collection itself, will be the same for each of the strategies:

"We will recruit 39 professional recruiters. Each recruiter is assigned at random to the "written" or "spoken" condition. [+some detailed description of the materials.] After reading/hearing the job pitch, the recruiter scores the candidate on three categories. The average score will be designated the intellect rating."

We should note in passing that the sample size in this example is quite small, meaning that this study only has adequate power to detect fairly large effects. For example, if the primary analysis was to be an independent-samples Student's *t* test (with a 2-sided test), and if the assumptions of the Student's *t* test were satisfied, then this study would have just 33% power to detect a "medium" effect size of $d = 0.5$. The power to detect an effect of this magnitude would be lower again for the other analysis methods specified below, even if the *t* test assumptions were met, and potentially even lower again in the presence of violations of the assumptions of these methods (depending on the nature of those assumption violations). In short, the data are useful for the purposes of illustration, but the sample size is not one that we would typically recommend. Please note that this is no criticism towards Schroeder and Epley (2015), as they base their conclusions on four separate experiments, three of which have a sample size much larger than 39.

For the sake of brevity, we have not included individual power analyses for each of the four separate strategies we illustrate.

Strategy 1, the Decision Tree

To compare the intellect ratings of both groups, we could apply a Student's *t* test. However, this test assumes (i) normality of the dependent variable within each of the two populations, (ii) equal variances between populations. We will therefore apply the following decision tree:

1. First, we apply the Shapiro Wilk for normality test on the deviations from the group average and denote the resulting *p*-value by p_{SW}.
    a. If $p_{SW} < 0.05$ we deem the normality assumption breached and we will apply the non-parametric Mann-Whitney/Wilcoxon test.

   b. If $p_{SW} > 0.05$, there is no evidence to reject the assumption of normality and we proceed with:

2. Second, we apply the *F* test for equality of variances and denote the resulting *p*-value by p_F.
   a. If $p_F < 0.05$ we deem this assumption violated and we will apply the Welch *t* test.
   b. If $p_F > 0.05$, we will apply Student's *t* test.

For the chosen test, we will report the *p* value of the two-sided comparison.

Note that this decision tree is just one way to check the assumptions. We chose the Shapiro-Wilks procedure after the recommendation by Razali and Wah (2011), but could have employed other normality tests instead. A similar argument holds for the *F* test for equality of variances. Furthermore, the order of testing both assumptions matters, but is arbitrary.

Strategy 2, Robust Primary Analysis Strategy

To compare the intellect ratings of both groups, we will use a bootstrapped version of Yuen's (1974) two-sample trimmed *t* test. This test is an alternative to the independent samples Student's *t* test designed for situations where there are both unequal variances across groups *and* non-normality within groups. The test is particularly useful with very long-tailed error distributions. To apply this test, we use the function yuenbt from the WRS2 package (Mair & Wilcox, 2018) in the statistical software R (R Core Team, 2018). We trim 20% of the data, thus removing the 10% smallest and 10% largest observations. This way, the influential effect of outliers is diminished. We employ 1,000 bootstrap samples, which is sufficient for accurate results (Mair & Wilcox, 2018). Before bootstrapping we set the random number generator to seed 1. We will report the *p* value of this test as well as the 95% confidence interval for the trimmed mean.

Strategy 3, Robustness Analysis

To compare the intellect rating of both groups, we will perform the following tests, each based on a different set of assumptions: (i) Student's *t* test, (ii) Welch *t* test, (iii) the Mann-Whitney/Wilcoxon test, (iv) Yuen's test for the 20% trimmed means, (v) the bootstrap version of test (iv) with 1,000 bootstrap samples and a seed set to 1. This last test mimics that used in Strategy 2 above.

Strategy 4, Exploratory Assumption Checks

To compare the intellect rating of both groups, we perform Student's *t* test. We also provide exploratory checks of the validity of the distributional assumptions underlying this test.

Example: Results for Each Strategy

In this case we obviously already have the data, so we can see how the four preregistrations work out:

Strategy 1, decision tree. The Shapiro-Wilk test was non-significant ($p = .124$), and so was the two-sided *F* test for equality of variances ($p = .458$). We therefore conducted Student's *t* test, and this test found a significant difference between both groups, $t(37) = -3.525$, $p = .001$.

Strategy 2, robust primary analysis strategy. The bootstrap version of the Yuen test provided a *p* value of .002 and a 95% confidence interval for the trimmed mean difference [-3.323, -0.698].

Strategy 3, robustness analysis. All five tests yielded a significant difference (at $\alpha = 0.05$) in favour of the audio group, with the following test results. Student's *t*: $t(37) = -3.525$, $p = .001$; Welch's *t*: $t(33.441) = -3.478$, $p = .001$; Wilcoxon test: $W = 84.5$, $p = .003$; Yuen's test: trimmed mean difference -2.010, $p = .004$, bootstrapped Yuen's test: trimmed mean difference -2.010, $p = .002$. Thus, we conclude that applicants with a spoken job pitch receive higher ratings than those with a written job pitch.

Strategy 4, exploratory assumption checks. Student's *t* test indicated that the audio-group performed significantly better than the written-group, $t(37) = -3.525$, $p = .001$. The Shapiro-Wilk test for normality provided no significant evidence for non-normality ($W = 0.966$, $p = .124$) and the *F* test for equality of variances did not provide significant evidence for violation of this assumption, $F(17, 20) = 1.411$, $p = .458$.

Example: Summary

It will be clear that all four strategies have their own benefits and drawbacks. A clear benefit of Strategy 2, for instance, is that the results can be reported in a very condensed form. A drawback, however, is that fewer people are familiar with the

Yuen test compared to more conventional tests. In this example, we deliberately kept the methodology as simple as possible with a single test for a two-group comparison. The complexity of some strategies will grow beyond feasible limits when the research questions are more complex.

Summary: Strategies for Distributional Assumption Checks

In the subsections above, we have outlined four strategies for dealing with distributional assumptions in preregistered research. Obviously, however, we have not considered all possible approaches that researchers could take to dealing with distributional issues in a preregistration. Of the four strategies we have discussed, strategy 1 (a decision tree) seems the least desirable on several counts, although it may be useful in some research contexts—particularly when applied by researchers with particularly strong statistical expertise. Strategy 3 (preregistering robustness analyses) is probably the most comprehensive and sophisticated way of dealing with the uncertainty arising from distributional assumptions. It is nevertheless a relatively challenging strategy to adequately specify in a preregistration (and to perform a power analysis for), and the strategy that would result in the most verbose write-up of the results. On the other hand, strategy 4 would be the easiest to specify in a preregistration, and may be useful for student research, or for researchers new to preregistration. Finally, Strategy 2 represents something of a compromise between difficulty level and level of sophistication, would produce a very concise write-up of results, and could also be applied in conjunction with any of the three other strategies.

Additional Tactics for Dealing with Distributional Assumptions

The four strategies listed above are, to at least some degree, competing strategies. On the other hand, there are two additional tactics that may be useful in virtually any preregistered research project, and that can be employed in conjunction with whichever of the four strategies above is selected.

Tactic 1: Clearly and Accurately Describe Distributional Assumptions

The first of these tactics is to transparently and accurately describe the distributional assumptions of the statistical analyses employed, and then acknowledge in the limitations section of the discussion that any uncertainty with respect to the validity of these assumptions adds to the uncertainty surrounding the substantive conclusions. A description of the assumptions of various statistical analyses is beyond the scope of this article, but some useful sources include Gelman and Hill (2007), Williams et al. (2013), and Casella and Berger (2002). We note in passing that is not uncommon for distributional assumptions to be described incorrectly in resources aimed at psychologists; it can often be useful to look to more rigorous sources written for statisticians.

Tactic 2: Open Data

The second tactic which we suggest applying in virtually any study—subject to any restrictions necessary to safeguard the privacy of the participants, or necessary for legal reasons—is to make a de-identified copy of the raw data openly accessible for readers and reviewers (see Houtkoop et al., 2018; Munafò et al., 2017; Nosek et al., 2015). Although the authors of any given piece of research will always have the primary responsibility for conducting and reporting data analyses that address uncertainty arising due to distributional assumptions, sharing open data nevertheless helps to ensure that others who might wish to apply a different approach to checking distributional assumptions—or a different approach to the primary analyses—are able to check the robustness of the findings to such alternative approaches. Along with the raw data it is useful to post the programming code or syntax necessary to process the data and apply the analyses reported in a given article. The Open Science Framework (https://osf.io) is a particularly useful venue for posting data and analysis syntax, but other options are available—including posting supplementary materials along with the article on a journal's website.

When preregistering a study that will use an open data policy, it is important to consider how this will be signalled to participants and in any institutional review board/ethics committee application (see

Meyer, 2018). In some parts of the world, institutional review boards may expect that data will be kept entirely confidential to the research team, and boilerplate information sheet and consent materials may encode this expectation. It is thus crucial to plan an open data policy from the beginning of a project, rather than leaving any decisions about data sharing to after data has been collected (at which point the researchers may find themselves inadvertently locked in to a restrictive data sharing policy).

Templates for Preregistrations

In order to assist researchers prepare preregistrations that pre-emptively deal with the prospect of distributional assumption violations, we suggest that preregistration templates include prompts leading researchers to consider employ some of the strategies described above. The specific prompts we would suggest would be:

> What are the distributional assumptions of the statistical analyses you will be applying?
>
> How have you accounted for the possibility of violations of these distributional assumptions? (Some options include selecting analysis methods that make assumptions that are as conservative as possible; preregistering robustness analyses which test the robustness of your findings to analysis strategies that make different assumptions; and/or pre-specifying a single primary analysis strategy, but noting that you will also report an exploratory investigation of the validity of distributional assumptions).

Conclusion

Preregistration is a valuable strategy in confirmatory research projects, but it does come with challenges. One of those challenges is the need to make decisions about how distributional assumption violations will be dealt with before examining the data itself. In this article, we have examined several strategies that researchers could adopt for addressing distributional assumptions in preregistered research. While preregistering "decision trees" for changing the primary analysis depending on the presence of particular assumption breaches has several problems, preregistering analyses that make weaker distributional assumptions, or preregistering robustness analyses, may be more useful approaches. Alternatively, students and other researchers new to preregistration may find it useful to preregister a simple primary analysis strategy—and stick with that regardless of the characteristics of the data—but also conduct and report an exploratory *post hoc* analysis of the validity of the distributional assumptions made. We recommend that researchers use the guidance above to select the strategy—or combination of strategies—that is most appropriate for their given context. Finally, we suggest that transparently and accurately communicating the assumptions of the analyses employed, and making raw data openly available, can be useful tactics for ensuring that readers and reviewers have sufficient information available to reach an informed judgment about the impact of any distributional problems on the validity of a study's conclusions.

References

Albers, W., Boon, P. C., & Kallenberg, W. C. M. (2000). Size and power of pretest procedures. *The Annals of Statistics*, *28*(1), 195–214. https://doi.org/10.1214/aos/1016120369

Azzalini, A. (1985). A class of distributions which includes the normal ones. *Scandinavian Journal of Statistics*, *12*(2), 171–178.

Bancroft, T. A. (1964). Analysis and inference for incompletely specified models involving the use of preliminary test(s) of significance. *Biometrics*, *20*(3), 427–442. https://doi.org/10.2307/2528486

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., … Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software*, *76*(1). https://doi.org/10.18637/jss.v076.i01

Casella, G., & Berger, R. L. (2002). *Statistical inference* (2nd ed.). Pacific Grove, CA: Duxbury.

Delacre, M., Lakens, D., & Leys, C. (2017). Why psychologists should by default use Welch's *t*-test instead of Student's *t*-test. *International Review of Social Psychology*, 30(1), 92–101. https://doi.org/10.5334/irsp.82

Ernst, A. F., & Albers, C. J. (2017). Regression assumptions in clinical psychology research practice—a systematic review of common misconceptions. *PeerJ*, 5. https://doi.org/10.7717/peerj.3323

Etz, A., & Vandekerckhove, J. (2018). Introduction to Bayesian inference for psychology. *Psychonomic Bulletin & Review*, 25(1), 5–34. https://doi.org/10.3758/s13423-017-1262-3

Fagerland, M. W., & Sandvik, L. (2009). The Wilcoxon–Mann–Whitney test under scrutiny. *Statistics in Medicine*, 28(10), 1487–1497. https://doi.org/10.1002/sim.3561

Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, 41(4), 1149–1160. https://doi.org/10.3758/BRM.41.4.1149

Ferguson, C. J., & Heene, M. (2012). A vast graveyard of undead theories: Publication bias and psychological science's aversion to the null. *Perspectives on Psychological Science*, 7(6), 555–561. https://doi.org/10.1177/1745691612459059

Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models*. Cambridge, United Kingdom: Cambridge University Press.

Gelman, A., & Loken, E. (2014). The statistical crisis in science. *American Scientist*, 102(6), 460–465. https://doi.org/10.1511/2014.111.460

Houtkoop, B. L., Chambers, C., Macleod, M., Bishop, D. V. M., Nichols, T. E., & Wagenmakers, E.-J. (2018). Data sharing in psychology: A survey on barriers and preconditions. *Advances in Methods and Practices in Psychological Science*, 1(1), 70–85. https://doi.org/10.1177/2515245917751886

Kruschke, J. K., & Liddell, T. M. (2018). Bayesian data analysis for newcomers. *Psychonomic Bulletin & Review*, 25(1), 155–177. https://doi.org/10.3758/s13423-017-1272-1

Liu, R. Y. (1988). Bootstrap procedures under some non-iid models. *The Annals of Statistics*, 16(4), 1696–1708. https://doi.org/10.1214/aos/1176351062

Lumley, T., Diehr, P., Emerson, S., & Chen, L. (2002). The importance of the normality assumption in large public health data sets. *Annual Review of Public Health*, 23, 151–169. https://doi.org/annurev.publhealth.23.100901.140546

Mair, P., & Wilcox, R. (2018). WRS2: A collection of robust statistical methods (Version 0.10-0). Retrieved from https://CRAN.R-project.org/package=WRS2

McIntyre, K. P. (2016). Do spoken or written words better express intelligence? Retrieved August 3, 2018, from Open Stats Lab website: https://sites.trinity.edu/osl/data-sets-and-activities/t-test-activities

Meuleman, B., Loosveldt, G., & Emonds, V. (2015). Regression analysis: Assumptions and diagnostics. In H. Best & C. Wolf (Eds.), *Regression analysis and causal inference* (pp. 83–110). London, United Kingdom: Sage.

Meyer, M. N. (2018). Practical tips for ethical data sharing. *Advances in Methods and Practices in Psychological Science*, 1(1), 131–144. https://doi.org/10.1177/2515245917747656

Munafò, M. R., Nosek, B. A., Bishop, D. V. M., Button, K. S., Chambers, C. D., Sert, N. P. du, … Ioannidis, J. P. A. (2017). A manifesto for reproducible science. *Nature Human Behaviour*, 1. https://doi.org/10.1038/s41562-016-0021

Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S. J., … Yarkoni, T. (2015). Promoting an open research culture. *Science*, 348(6242), 1422–1425. https://doi.org/10.1126/science.aab2374

Quintana, D. S. (2015). From pre-registration to publication: A non-technical primer for conducting a meta-analysis to synthesize correlational data. *Frontiers in Psychology*, 6. https://doi.org/10.3389/fpsyg.2015.01549

R Core Team. (2018). *R: A language and environment for statistical computing*. Retrieved from http://www.R-project.org/

Razali, N. M., & Wah, Y. B. (2011). Power comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling tests. *Journal of Statistical Modeling and Analytics*, *2*(1), 21–33.

Saltelli, A., Tarantola, S., Campolongo, F., & Ratto, M. (2004). *Sensitivity analysis in practice: A guide to assessing scientific models*. Chichester, United Kingdom: Wiley.

Schroeder, J., & Epley, N. (2015). The sound of intellect: speech reveals a thoughtful mind, increasing a job candidate's appeal. *Psychological Science*, *26*(6), 877–891. https://doi.org/10.1177/0956797615572906

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*(11), 1359–1366. https://doi.org/10.1177/0956797611417632

Steegen, S., Tuerlinckx, F., Gelman, A., & Vanpaemel, W. (2016). Increasing transparency through a multiverse analysis. *Perspectives on Psychological Science*, *11*(5), 702–712. https://doi.org/10.1177/1745691616658637

van 't Veer, A. E., & Giner-Sorolla, R. (2016). Pre-registration in social psychology—A discussion and suggested template. *Journal of Experimental Social Psychology*, *67*, 2–12. https://doi.org/10.1016/j.jesp.2016.03.004

Wagenmakers, E.-J., Wetzels, R., Borsboom, D., van der Maas, H. L. J., & Kievit, R. A. (2012). An agenda for purely confirmatory research. *Perspectives on Psychological Science*, *7*(6), 632–638. https://doi.org/10.1177/1745691612463078

Wells, C. S., & Hintze, J. M. (2007). Dealing with assumptions underlying statistical tests. *Psychology in the Schools*, *44*(5), 495–502. https://doi.org/10.1002/pits.20241

Westfall, J., & Yarkoni, T. (2016). Statistically controlling for confounding constructs is harder than you think. *PLOS ONE*, *11*(3), e0152719. https://doi.org/10.1371/journal.pone.0152719

White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica: Journal of the Econometric Society*, *48*(4), 817–838. https://doi.org/10.2307/1912934

Wilcox, R. (2012). *Introduction to robust estimation and hypothesis testing* (3rd ed.). Waltham, MA: Academic Press.

Wilkinson, L., & Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, *54*(8), 594–604. https://doi.org/10.1037/0003-066X.54.8.594

Williams, M. N., Grajales, C. A. G., & Kurkiewicz, D. (2013). Assumptions of multiple regression: Correcting two misconceptions. *Practical Assessment, Research & Evaluation*, *18*(11). Retrieved from http://www.pareonline.net/getvn.asp?v=18&n=11

Yuen, K. K. (1974). The two-sample trimmed t for unequal population variances. *Biometrika*, *61*(1), 165–170. https://doi.org/10.1093/biomet/61.1.165

Zimmerman, D. W. (2004). A note on preliminary tests of equality of variances. *British Journal of Mathematical and Statistical Psychology*, *57*(1), 173–181. https://doi.org/10.1348/000711004849222